



Item Analysis, Reliability Estimates, Percentile Ranks and Stanine Scores Among University Students in Ghana

Abraham Yeboah

Department of Interdisciplinary Studies, Akenten Appiah-Menka
University of Skills Training and Entrepreneurial
Development (AAMUSTED), Kumasi, Ghana
ORCID: 0000-0001-9310-4367

Abraham Gyamfi

Department of Education
Wesley College of Education, Kumasi, Ghana

Dominic Kaku Wintson

Renmin University of China

Akosua Dwomoh Prempeh

Department of Cognitive Science
Technische Universität, Kaiserslautern, Germany

ABSTRACT

The objective of the study was to perform item analysis on a forty (40) multiple choice test items among students of the University of Cape Coast in Ghana. The study further sought to estimate the reliability of the results and compute the percentile ranks and stanine scores. The results showed that, regarding the two statistics; difficulty (*P*) and discrimination (*D*) indices, it was observed that twelve (12) items functioned properly. Five items, although were easy, discriminated well. Nine (9) items were either easy or difficult and had low discrimination indices. It was found that nine (9) items had acceptable difficulty level but did not discriminate well. The results revealed that two (2) items had negative discrimination indices and two items were too easy and did not discriminate. The estimate of reliability; Spearman Brown prophecy (r_{xy}) and Cronbach alpha (α) were found to be 0.63 and 0.78 respectively. It was found that examinees with scores of 33 had 99% of the norm group performing below them while examinees with scores of 26 had 70% of the norm group falling below them. Examinees with scores of 32 and 33 were found to be in stanine nine, those with scores of 23 to 24 were in stanine five whilst examinees with scores of 4 to 16 were in stanine one. Test results are essential for the determination of learner's future endeavours. It is recommended that test item developers (teachers) in schools, colleges and universities adhere to principles of test construction, administration and interpretation so as to enhance dependability of learners' test results.

Keywords: discrimination, difficulty, reliability, examinees.

INTRODUCTION

There are many ways of inferring sample of behaviours and traits from either a group of individuals or a particular person. The tools used for measuring the traits include paper-and-pencil tests, portfolios, projects, oral presentation, observation, learning journals, learning contracts, and interview (Knight & Yorke, 2003). This write up employed paper and pencil test to determine the extent to which university students had mastered the content and behavioural outcomes stipulated in Educational Statistics course in the University of Cape Coast. The test consisted of 40 multiple-choice items which were based on the content in the syllabus for which students had been taught thereby ensuring content related evidence of validity of the test.

Construction of Test Items

The test was planned with regard to the purpose of the test before constructing the items. To achieve this, a table of specification was prepared for the test items. The test items were developed in adherence to the Benjamin Bloom's (1956) taxonomy of educational objectives. Quality classroom-based assessment means adherence to standard procedures for test construction. Every classroom teacher is expected to possess and apply requisite skills in construction of good items for class assessments. Based on the proposition of Tamakloe, Atta and Amedahe (1996) and Ester (2004), we considered the following steps when constructing the test items by defining the purpose of the test, determining the item format to use, determining what is to be tested, writing the individual items, reviewing the items, preparing the scoring key, writing directions, and evaluating the test.

Similarly, to ensure general applicability and the superior qualities of multiple-choice test items, we made sure that we formulated items which clearly stated the problem, identified plausible alternatives, and removed irrelevant clues to keys or answers. Miller, Linn and Gronlund (2009) suggested the following principles for writing multiple-choice items:

1. The stem of the item should be meaningful by itself and should present a definite problem.
2. The item stem should include as much of the item as possible and should be free of irrelevant material. This will increase the probability of a clearly stated problem in the stem and will reduce the reading time required.
3. Use a negatively stated stem only when significant learning outcomes require it.
4. All the alternatives should be grammatically consistent with the stem of the item.
5. An item should contain only one correct or clearly best answer.

Administration of Test Items

The rationale for test administration was to ensure fairness to all examinees to demonstrate their understanding of what had been taught. Oduro-Okyireh (2008) reported that the guiding principle in test administration is to provide all examinees with a fair chance to demonstrate their achievement on what is being measured. To ensure effective test administration procedures, we were guided by the findings of Tyler (1991) which are useful in actual practice in test administration of any testing programmes. To achieve this, we ensured that the following as pointed out by Tyler (1991) where adhered to:

1. Select the tests carefully, preferably in co-operation with a college assessment committee.
2. Order the tests well in advance of the date on which they are to be used.
3. Avoid overemphasis on the tests.

4. Seat the students in alternate chairs, if possible.
5. Make announcements slowly and clearly in a voice that is loud enough to be heard throughout the room.
6. Have the blanks on the front of the booklets filled out. Be sure to announce the date, how names are to be written, and other items that may need clarification.
7. Time the examination with extreme care, using a watch or a wall clock
8. Move about the room occasionally to see that students are working on the right part of the test, but do not stand gazing over a student's shoulder and do not constantly move nervously from student to student.
9. Stop the test immediately when the time is up and collect the answer booklets.
10. As soon as a certain test has been given, have all the students taking the test turn in their booklets promptly. Alphabetize and check the test papers against the class list.
11. Except in cases of illness, see that all absentees make up the test. This necessity will cause the tutor much trouble and worry, but it is unavoidable, for complete data are essential if the results of the test are to be used successfully in either teaching or guidance.

Scoring of Test Items

Test items scoring is critical in assessment. It provides examinees with outcomes of their examination for decision making. Conventionally, there are two methods for scoring multiple choice items: Number right scoring and Right minus wrong scoring. Number right scoring is where correct answers are scored with a positive value, incorrect answers and absent or omitted answers with a value of zero whilst Rights minus wrongs scoring (Betts, Elder, Hartley & Trueman, 2009) indicated that students acknowledge they will lose marks for incorrect answers and therefore are very circumspect responding to test items. In this write-up, the number right scoring method was adopted. With number right method of scoring, students without the ability to solve a particular item, gain marks by guessing (Kubinger, Holocher-Ertl, Reif, Hohensinns & Frebort, 2010). Test designers can as such not distinguish between correct answers based on knowledge mastery versus those based on a guess (Bar-Hillel, Budescu & Attali, 2005). The results of the testees or examinees are presented in Table 2 in a descending order of performance. The results showed that the maximum score was 33 and the minimum was 4. The mean score was 22.26 and the standard deviation score also 6.23. The mean score of 22.26 indicated that examinees performed fairly well and standard deviation of 6.23 also implied that the scores were not widely spread.

Item Analysis

Item analysis refers to the process of collecting, summarizing and using information from students' responses to make decisions about each assessment task Nitko (1996). Tamakloe, Atta and Amedahe (1996) opined that item analysis is described as a post mortem evaluation and further explained that it is a group of procedures for assessing the quality of examination items.

The purpose of item analysis is to improve the quality of teaching and also examination as a whole. The analysis of students' results will point out teachers' strengths and weakness in terms of teaching and learning, therefore to ensure that, the right thing is done by the teacher. Again, item analysis helps one to identify items for retention, revision or removal. Fundamentally, the aim of item analysis is to improve items of a test for future used based on

indices of item difficulty and item discrimination as well as the extent of effective functioning of the distracters.

Particularly, two numerical indicators are often computed during item analysis. These are item difficulty (**P**) and item discrimination (**D**).

Item difficulty (**P**)

Amedahe and Asamoah-Gyimah (2003) reported item difficulty as the percentage of examinees who answer correctly each test item. It is calculated by dividing number of examinees who answer correctly the item by total number of examinees. Mathematically, $P = \frac{R}{T}$ where R = number of examinees who answer correctly the item and T= total number of examinees. The P index ranges from 0 when no examinee correctly answered the item to 1 when all examinees answered the item correctly. This indicates that the smaller the P index, the more difficult the item and the greater the P index, the less difficult the item. It should be noted that item difficulty is calculated for each item but the entire test. Table 1 shows the P of each of the 40 items with 40 examinees.

Table 1: Results of Item Difficulty Index for Examinees

Item No.	Total number of students who correctly answered the item (R)	Proportion Correct (P) = $\frac{R}{T}$
1	22	0.59
2	21	0.57
3	34	0.92
4	13	0.35
5	21	0.57
6	23	0.62
7	05	0.14
8	10	0.27
9	16	0.43
10	09	0.24
11	27	0.73
12	24	0.65
13	19	0.51
14	15	0.41
15	24	0.65
16	24	0.65
17	37	1.00
18	28	0.76
19	29	0.78
20	34	0.92
21	14	0.38
22	22	0.59
23	20	0.54
24	28	0.76

25	23	0.62
26	18	0.49
27	17	0.46
28	19	0.51
29	05	0.14
30	28	0.76
31	29	0.78
32	19	0.51
33	09	0.24
34	20	0.54
35	24	0.65
36	06	0.16
37	30	0.81
38	24	0.65
39	27	0.73
40	30	0.81

Table 1 shows the statistics of number of examinees who correctly answered items and the difficulty level for each item. Item difficulty (P) is the proportion of examinees who score an item correctly in relation to the number of examinees who attempted the item (Pido, 2012). Pido added that the smaller the P-value the more difficult the item and when the P-value is large the item is easy. The P- index ranges between 0 and 1. Generally the recommended item P-value ranges between 0.3 and 0.7 to maximize test information and differences among the examinees (Allen & Yen, 1979).

Based on the above criteria then twenty-two (22) of the items functioned well. These are items 1, 2, 4, 5, 6, 9, 12, 13, 14, 15, 16, 21, 22, 23, 25, 26, 27, 28, 32, 34, 35 and 38 with P- indices between 0.3 and 0.7. Six (6) of the items were found to be difficult (items 7, 8, 10, 29, 33 and 36) with P-indices less than 0.3 and twelve (12) of the items were easy (items 3, 11, 17, 18, 19, 20, 24, 30, 31, 37, 39 and 40) with P-indices greater than 0.7. Notwithstanding, one cannot use item difficulty only to determine the effectiveness of items therefore, the need to find out how items discriminate between examinees.

Item Discrimination

Based on the scores of the students, 12 examinees fell in the upper group, 13 in the middle group and 12 in the lower group. This was done by calculating 0.33(33%) of the 37 examinees for each group after the scores have been arranged from the highest to lowest as suggested by Tamakloe, Atta and Amedehe (1996). Table 2 presents the results of the examinees.

Table 2: Distribution of Results of Examinees in a Descending Order

Examinees	Scores	Level
01s	33	U
02	32	U
03	30	U
04	30	U
05	30	U
06	28	U
07	28	U
08	27	U
09	27	U
10	27	U
11	26	U
12	26	U
13	25	M
14	24	M
15	24	M
16	24	M
17	24	M
18	23	M
19	23	M
20	23	M
21	22	M
22	21	M
23	21	M
24	21	M
25	21	M
26	20	L
27	20	L
28	20	L
29	18	L
30	18	L
31	17	L
32	16	L
33	15	L
34	15	L
35	13	L
36	09	L
37	04	L

Tamakloe, Atta and Amedahe (1996) defined item discrimination as a measure of the degree to which an item discriminates between examinees with high performance and examinees with low performance. A discrimination index of 0.5 is of average discrimination power for standardized tests, a discrimination index of 0.4 or better is satisfactory. Items with discrimination index below 0.2 must either be discarded or rewritten (Tamakloe, Atta & Amedahe, 1996). A negative D-value indicates that more of the low achievers got the item correct than the high achievers. Such items are ambiguous and therefore, must be either

discarded or reviewed. Table 3 presents the results of 40 multiple-choice items discrimination index for the upper and the lower groups respectively.

Table 3: Results of Item Discrimination Index for Upper and Lower Groups

Item	R_U	$P_U = R_U/N_U$	R_L	$P_L = R_L/N_L$	$D = P_U - P_L$
1	10	0.83	5	0.42	0.42
2	10	0.83	3	0.25	0.58
3	12	1.00	10	0.83	0.17
4	7	0.58	3	0.25	0.33
5	7	0.58	5	0.42	0.16
6	9	0.75	5	0.42	0.33
7	0	0.00	2	0.17	-0.17
8	6	0.50	2	0.17	0.33
9	8	0.67	2	0.17	0.50
10	3	0.25	4	0.33	-0.08
11	11	0.92	5	0.42	0.50
12	8	0.67	6	0.50	0.17
13	9	0.75	5	0.42	0.33
14	7	0.58	4	0.33	0.25
15	11	0.92	7	0.58	0.34
16	9	0.75	5	0.42	0.33
17	12	1.00	12	1.00	0.00
18	10	0.83	6	0.50	0.33
19	10	0.83	11	0.92	-0.09
20	12	1.00	10	0.83	0.17
21	7	0.58	5	0.42	0.16
22	11	0.92	4	0.33	0.59
23	10	0.83	5	0.42	0.41
24	12	1.00	7	0.58	0.42
25	11	0.92	5	0.42	0.50
26	10	0.83	3	0.25	0.58
27	8	0.67	3	0.25	0.42
28	11	0.92	2	0.17	0.75
29	2	0.17	0	0.00	0.17
30	10	0.83	4	0.33	0.50
31	11	0.92	7	0.58	0.34
32	7	0.58	3	0.25	0.33
33	5	0.42	0	0.00	0.42
34	10	0.83	2	0.17	0.66
35	10	0.83	5	0.42	0.41
36	2	0.17	2	0.17	0.00
37	12	1.00	6	0.50	0.50
38	11	0.92	6	0.50	0.42

39	12	1.00	7	0.58	0.42
40	12	1.00	8	0.67	0.33

R_U = number of students in the upper group answering the item correctly,

R_L = number of students in the lower group answering the item correctly

N_U = number of students in upper group. In this case 14

N_L = number of students in lower group. In this case 11

P_U = proportion of students in upper group answering the item correctly

P_L = proportion of students in lower group answering the item correctly

D = the item discrimination index

Based on the above information, eighteen (18) items discriminated well. That is their D-indices were 0.4 or greater. The items were 1, 2, 9, 11, 22, 23, 24, 25, 26, 27, 28, 30, 33, 34, 35, 37, 38 and 39. Eleven (11) items indices were between 0.2 and 0.39 indicating that items discriminated satisfactorily. The items are 4, 6, 8, 13, 14, 15, 16, 18, 31, 32 and 40 and six (6) items had low discriminating indices ($D < 0.2$). These items are 3, 5, 12, 20, 21 and 29. Three (3) items, 7, 10 and 19 had negative indices and two (2) items 17 and 36 also had 0.0 indices meaning these items did not discriminate between high and low achievers.

Regarding the two statistics above, (difficulty and discrimination indices) it could be seen that twelve (12) items (1, 2, 9, 22, 23, 25, 26, 27, 28, 34, 35 and 38) functioned properly meeting the acceptable levels of both P and D. Items 11, 24, 30, 37 and 39 although was easy, discriminated so well therefore require little revision. Items 3, 8, 18, 20, 21, 29, 31, 33 and 40 were either easy or difficult and have low discrimination indices. Items 4, 5, 6, 12, 13, 14, 15, 16 and 21 had acceptable difficulty level but are not discriminating well therefore require major revision as well. For items 7, 10 and 19, which had negative discrimination, the key must be rechecked or must be replaced. Items 17 and 36, must be discarded because they were too easy and did not discriminate.

Analysis of Distracters

Items' quality essentially depends partly on the effective functioning of the selected distracters. The real functioning's of the distracters is identified by logical check of the options and how they are selected by the examinees. Most importantly, every distracter must attract, at least one examinee. In general, a good distracter must attract more examinees from the lower group than the high Group (Tamakloe, Atta & Amedahe, 1996). Therefore, the distracters of some sampled items; 1, 5, 10, 15, 20, 25, 30, 35 and 40 are discussed below.

Item 1				
Option	U	M	L	Total
A	1	3	2	6
B*	10	6	5	21
C	1	2	5	8
D	0	2	0	2
Total	12	13	12	37

From the Table above, it is clearly seen that distracter D is not effective because it did not attract the low achievers therefore, this distracter must be replaced.

Item 5				
Option	U	M	L	Total
A*	7	8	5	20
B	2	3	2	7
C	1	0	1	2
D	2	2	4	8
Total	12	13	12	37

Item 5 functioned effectively because all distracters attracted low achievers. The key also functioned well because it attracted more high achievers than low achievers.

Item 10				
Option	U	M	L	Total
A	5	4	2	11
B	0	3	5	8
C*	3	2	4	9
D	4	4	1	9
Total	12	13	12	37

Item 10 is faulty because more low achievers responded well than high achievers. It could be that either the item is ambiguous or the key has a problem. Item 10 therefore needs re-examination. But all distracters functioned effectively because it attracted low achievers.

Item 15				
Option	U	M	L	Total
A	1	2	3	6
B	0	1	0	1
C	0	2	2	4
D*	11	8	7	26
Total	12	13	12	37

From the Table, distracter B did not function effectively. This is because it did not attract any low achiever. Distracter B needs replacement. However, the key functioned effectively because more high achievers selected than low achievers.

Item 20				
Option	U	M	L	Total
A	0	1	1	2
B*	12	12	10	34
C	0	0	1	1
D	0	0	0	0
Total	12	13	12	37

It can be observed that distracter D needs replacement. This is due to the fact that it did not attract any low achiever. The key functioned well because more high achievers selected than low achievers.

Item 25				
Option	U	M	L	Total
A*	11	10	5	26
B	1	1	2	4
C	0	1	2	3
D	0	1	3	4
Total	12	13	12	37

Item 25 had good distracters. All distracters attracted lower achievers. Also, the key functioned effectively because more high achievers selected it than low achievers.

Item 30				
Option	U	M	L	Total
A*	10	12	4	26
B	0	1	3	4
C	1	0	3	4
D	1	0	2	3
Total	12	13	12	37

Item 30 is good because all distracters attracted low achieving examinees. Moreover, the key functioned well because more high achievers selected it than low achievers.

Item 35				
Option	U	M	L	Total
A	0	0	3	3
B	2	3	3	8
C*	10	9	5	24
D	0	1	1	2
Total	12	13	12	37

Item 35 is good because all distracters attracted low achievers. Additionally, the key functioned well because more high achievers selected it than low achievers. Hence the results of item 35 is reliable.

Item 40				
Option	U	M	L	Total
A	0	1	1	2
B	0	0	1	1
C	0	1	2	3
D*	12	11	8	31
Total	12	13	12	37

Item 40 functioned effectively because all distracters attracted low achievers. The key also functioned well because it attracted more high achievers than low achievers.

RELIABILITY

The issue of consistent of examinees results is very crucial as far as assessment is concerned. Etsey (2012) assert that reliability is the degree of consistency of assessment results. Amedehe (2015) also states that reliability is a proportion and gives an indication of how much errors are in a set of scores. That is reliability is related to errors in scores. In estimating the reliability coefficient, the Split-half and Cronbach's alpha reliability coefficients were used. Table 4 shows the item variance of performance on each item.

Table 4: Results of Item Difficulty Index for Examinees

Item No.	Total number of students who correctly answered the item (R)	Proportion Correct (p) = $\frac{R}{T}$	q= 1-p	Pq
1	22	0.59	0.41	0.24
2	21	0.57	0.43	0.25
3	34	0.92	0.08	0.07
4	13	0.35	0.65	0.23
5	21	0.57	0.43	0.25
6	23	0.62	0.38	0.24
7	05	0.14	0.86	0.12
8	10	0.27	0.73	0.19
9	16	0.43	0.57	0.25
10	09	0.24	0.76	0.18
11	27	0.73	0.27	0.19
12	24	0.65	0.35	0.23
13	19	0.51	0.49	0.25

14	15	0.41	0.59	0.24
15	24	0.65	0.35	0.23
16	24	0.65	0.35	0.23
17	37	1.00	0.00	0.00
18	28	0.76	0.24	0.18
19	29	0.78	0.22	0.17
20	34	0.92	0.08	0.07
21	14	0.38	0.62	0.24
22	22	0.59	0.41	0.24
23	20	0.54	0.46	0.25
24	28	0.76	0.24	0.18
25	23	0.62	0.38	0.24
26	18	0.49	0.51	0.25
27	17	0.46	0.54	0.25
28	19	0.51	0.49	0.25
29	05	0.14	0.86	0.25
30	28	0.76	0.24	0.18
31	29	0.78	0.22	0.17
32	19	0.51	0.49	0.25
33	09	0.24	0.76	0.18
34	20	0.54	0.46	0.25
35	24	0.65	0.35	0.23
36	06	0.16	0.84	0.13
37	30	0.81	0.19	0.15
38	24	0.65	0.35	0.23
39	27	0.73	0.27	0.19
40	30	0.81	0.19	0.15
				$\Sigma=8.07$

Split-half Reliability

To ensure effective use of this reliability method, test items were divided into two parts. The first part covered items 1 to 20 and the second part covered items 21 to 40. Each half of the test were scored for each examinee. The score for each half of the test for each examinee was entered into the Statistical Package for Social Sciences (IBM, SPSS version 21). Pearson's correlation coefficient was run for the two halves and a coefficient value of 0.46 was obtained.

Table 5: Distribution of Results of First Half-Second Half of Examinees Scores

Examinees	First half (x)	Second half (y)
1	5	9
2	6	11
3	9	9
4	4	6
5	9	11
6	11	8
7	11	3
8	12	9
9	9	12
10	15	1
11	13	8
12	3	2
13	9	12
14	12	9
15	14	12
16	15	10
17	9	13
18	13	12
19	11	11
20	12	13
21	12	10
22	10	14
23	13	11
24	13	13
25	10	14
26	15	18
27	14	17
28	15	14
29	16	18
30	12	15
31	14	12
32	13	15
33	13	15
34	12	16
35	16	12
36	15	16
37	16	15

The Spearman-Brown prophecy formula which is often used to obtain the reliability coefficient for first half-second half was used to compute for the reliability. Etsey (2004) indicated that when using Spearman-Brown prophecy formula, test is divided into two halves for scoring. He

added that two scores for each student are correlated to obtain the estimate of reliability. The Spearman-Brown prophecy formula is presented below.

$$r_{xy} = \frac{2r}{1+r}$$

$$r_{xy} = \frac{2(0.46)}{1+0.46}$$

$$r_{xy} = \underline{0.63}$$

The reliability estimate obtained from the Spearman Brown prophecy formula is 0.63 or 63%. This clearly means that the error margin of examinees scores for the two halves of test is 0.37 or 37%. This implies that 63% of the examinees score could be used for decision making based on the purpose of the test.

Cronbach's Alpha Reliability

The Cronbach alpha is another method of estimating internal consistency. Cronbach alpha is the average split-half correlation based on all possible divisions of a test into two parts. This internal consistency is used when test items are scored pass-fail or when more than one point is awarded for a correct response (Salvia & Yesseldyke, 2001). Formula for Cronbach alpha is presented below.

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k p_i q_i}{s_x^2} \right)$$

$$\alpha = \frac{40}{40-1} \left(1 - \frac{8.07}{6.23^2} \right)$$

$$\alpha = 1.03 (1 - 0.21)$$

$$\alpha = \underline{0.78}$$

The reliability estimate obtained from the Cronbach alpha formula is 0.78 or 78%. This clearly means that the error margin of examinees scores for the two halves of test is 0.22 or 22%.

Percentile rank

Percentile ranks tells the percentage of the students in a norm group that have scored lower than the raw score in question. Table 7 presents the percentile ranks of scores obtained from examinees.

Percentile rank is computed by:

$$PR = \frac{1}{2} \left[\frac{\text{Number of persons having the score} + \text{Number of persons below the score}}{\text{Total number of persons}} \right] \times 100$$

Table 6: Percentile Ranks of Scores

Scores	Tally	Frequency	Cumulative frequency	Percentile rank
33	1	1	37	$\frac{1}{2} \frac{[1] + 36}{37} \times 100 = 99$
32	1	1	36	$\frac{1}{2} \frac{[1] + 35}{37} \times 100 = 96$
31		0	35	$\frac{1}{2} \frac{[0] + 35}{37} \times 100 = 95$
30	111	3	35	$\frac{1}{2} \frac{[3] + 32}{37} \times 100 = 87$
29		0	32	$\frac{1}{2} \frac{[0] + 32}{37} \times 100 = 86$
28	11	2	32	$\frac{1}{2} \frac{[2] + 30}{37} \times 100 = 84$
27	111	3	30	$\frac{1}{2} \frac{[3] + 27}{37} \times 100 = 73$
26	11	2	27	$\frac{1}{2} \frac{[2] + 25}{37} \times 100 = 70$
25	1	1	25	$\frac{1}{2} \frac{[1] + 24}{37} \times 100 = 66$
24	1111	4	24	$\frac{1}{2} \frac{[4] + 20}{37} \times 100 = 59$
23	111	3	20	$\frac{1}{2} \frac{[3] + 17}{37} \times 100 = 50$
22	1	1	17	$\frac{1}{2} \frac{[1] + 16}{37} \times 100 = 45$
21	1111	4	16	$\frac{1}{2} \frac{[4] + 12}{37} \times 100 = 38$
20	111	3	12	$\frac{1}{2} \frac{[3] + 9}{37} \times 100 = 28$
19		0	9	$\frac{1}{2} \frac{[0] + 9}{37} \times 100 = 24$
18	11	2	9	$\frac{1}{2} \frac{[2] + 7}{37} \times 100 = 22$
17	1	1	7	$\frac{1}{2} \frac{[1] + 6}{37} \times 100 = 18$
16	1	1	6	$\frac{1}{2} \frac{[1] + 5}{37} \times 100 = 15$
15	11	2	5	$\frac{1}{2} \frac{[2] + 3}{37} \times 100 = 11$
14		0	3	$\frac{1}{2} \frac{[0] + 3}{37} \times 100 = 8$
13	1	1	3	$\frac{1}{2} \frac{[1] + 2}{37} \times 100 = 7$
12		0	2	$\frac{1}{2} \frac{[0] + 2}{37} \times 100 = 5$
11		0	2	$\frac{1}{2} \frac{[0] + 2}{37} \times 100 = 5$
10		0	2	$\frac{1}{2} \frac{[0] + 2}{37} \times 100 = 5$
9	1	1	2	$\frac{1}{2} \frac{[1] + 1}{37} \times 100 = 4$

				37
8	0	1		$\frac{1}{2} \frac{[0] + 1}{37} \times 100 = 3$
7	0	1		$\frac{1}{2} \frac{[0] + 1}{37} \times 100 = 3$
6	0	1		$\frac{1}{2} \frac{[0] + 1}{37} \times 100 = 3$
5	0	1		$\frac{1}{2} \frac{[0] + 2}{37} \times 100 = 3$
4	1	1	1	$\frac{1}{2} \frac{[1] + 0}{37} \times 100 = 1$
Total	37			

From Table 6, an examinee with a score of 33 has 99% of the norm group performing below him/her. It was also observed that examinees with scores of 26 have 70% of the norm group falling below them.

Stanine

A stanine score tells the location of a raw score in a specific segment of a normal distribution. Stanine computations of raw scores for this write-up is presented below.

Stanine 1

$$^4/_{100} \times 37 = 1.48 \text{ or } 2$$

Stanine 2

$$^7/_{100} \times 37 = 2.59 \text{ or } 3$$

Stanine 3

$$^{12}/_{100} \times 37 = 4.44 \text{ or } 4$$

Stanine 4

$$^{17}/_{100} \times 37 = 6.29 \text{ or } 6$$

Stanine 5

$$^{20}/_{100} \times 37 = 7.40 \text{ or } 7$$

Stanine 6

$$^{17}/_{100} \times 37 = 6.29 \text{ or } 6$$

Stanine 7

$$^{12}/_{100} \times 37 = 4.44 \text{ or } 4$$

Stanine 8

$$^7/_{100} \times 37 = 2.59 \text{ or } 3$$

Stanine 9

$$^4/_{100} \times 37 = 1.48 \text{ or } 2$$

Table 7: Distribution of Stanine Results

Scores	Stanine	Actual number	Theoretical number
31-33	9	2	2
30	8	3	3
28-29	7	2	4
25-27	6	6	6
23-24	5	7	7
21-22	4	5	6
19-20	3	3	4
17-18	2	3	3
4-16	1	6	2

It is evident from Table 7 that 2 examinees with scores of 32 and 33 were found to be in stanine nine. It was also observed that 7 examinees with scores of 23 to 24 were actually in stanine five whilst 6 examinees with scores of 4 to 16 were also in stanine 1.

Implication for practice

1. The computations of item difficulty, discrimination and reliability using KR20 would enable teacher practitioners to determine whether assessment items truly functioned well or not, that is reliability and validity of items. Hence, it is practical for teachers to pay keen attention to item analysis.
2. In constructing multiple tests items, teachers will be encouraged to apply all the necessary rules in setting multiple tests items that are content based, more reliable and valid in nature.
3. Practicing teachers would be able to write distracters that are plausible in order to discriminate the higher achievers from the lower achievers in any given test.

Reference

- Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, California.
- Amedehe, F. K. (2015). *Reliability*. Oral conversation at Lectures. Cape Coast: Unpublished.
- Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple-choice test: a case study in irrationality. *Mind & Society*, 4, 3–12.
- Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education*, 34(1), 1–15.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krahwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. White Plains, NY: Longman.
- Etsey, Y. K. A. (2012). *Assessment in education*. Cape Coast: Unpublished
- Gyimah, A., & Amedehe, F (2003). *Introduction to measurement and evaluation*. Cape Coast: Unpublished
- Gyimah, A., & Amedehe, F (2003). *Introduction to measurement and evaluation*. Cape Coast: Unpublished
- Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, 18(1), 111–115.
- Nitko, A. J. (1996). *Educational assessment of students* (2nd Ed.). Englewood, NJ: Prentice Hall/Merrill Education.
- Nitko, A. J. (2004). *Educational Tests and Measurements* (3rd ed.). USA: Prentice-Hall, Inc

- Oduro-Okyireh, G. (2008). *Testing practices of senior secondary school teachers in the Ashanti Region of Ghana*. Unpublished master's thesis. University of Cape Coast, Cape Coast, Ghana.
- Pido, S. (2012). Comparison of item analysis results obtained using item response theory approaches. *Journal of Educational Assessment in Africa* 7 192-207.
- Salvia, J., & Ysseldyke, J. E. (2001). *Assessment in special and remedial education*. Boston: Houghton Mifflin.
- Tamakloe, E. K., & Amedahe, F. K. (1996). *Principles and Methods of Teaching*. Black Mask. Cantoment. Accra.
- Tyler, R. H. (1991). *Formal assessment in the classroom: The Ghana Education Service termly assessment plan*. Paper presented to the Quality Improvement in Primary Schools (QUIPS) Project. Funded by the USAID.